Semantic visualisation of gene expression information

Andy Taylor[1], Kenneth M^{c}Leod[1], and Albert Burger[1,2]

[1] Heriot-Watt University, Edinburgh, EH14 4AS, UK
ajt17 | kcm1 | a.g.burger @hw.ac.uk
WWW home page: http://www.macs.hw.ac.uk/bisel
[2] MRC Human Genetics Unit, Edinburgh, EH4 2XU, UK

Abstract. The Edinburgh Mouse Atlas of Gene Expression (EMAGE) publishes the results of gene expression experiments on the mouse embryo. Whilst this resource uses visual mechanisms to display the result of a single experiment, it currently provides no technique for the visual navigation of data. Ideally, a semantic visual navigation mechanism would exist. Our work focuses on trying to understand the requirements for such a mechanism. To this end, a prototype solution (based on sunburst visualisations) is being built. This paper presents the prototype and reports on the initial feedback from users.

Keywords: semantic visualisation, gene expression information, big data, usability

1 Introduction

Due to high throughput experiments and information technology, there are now many big data resources available for biologists. It is increasingly clear that users require assistance when navigating and searching this information. One way of providing support is through appropriate visualisation. Visualisation can be used to help users explore data or help users interpret information.

While such problems are widespread, this work focuses exclusively on a single use case. The Edinburgh Mouse Atlas of Gene Expression (EMAGE) [11] publishes online gene expression information for the developmental mouse. EMAGE provides a variety of tools to help users search the data; however, it does not enable users to visually navigate the data. This work aims to address this gap.

Although there are many different technologies and techniques that might be applicable (see Section 3), this work focuses on the application of sunburst visualisations. The goal is to develop a prototype application that will inform the creation of a real world solution. As such, the objective is not to create a powerful application that perfectly meets the needs of EMAGE's user community, but instead to understand the requirements of that group and learn how we may satisfy those needs in future.

This paper reports on the use case and motivation for this work, outlines the prototype currently under development and relates the initial feedback from potential end users.

Section 2 describes the use case in which this work is set, and reviews the existing visualisations that are employed within the use case. Section 3 considers related work before Section 4 describes how we chose the visualisation to focus on. Sunburst visualisations are discussed in Section 5. Subsequently, Section 6 reviews the customisation of Sunburst visualisations for use within the current use case. Section 7 features a discussion and conclusions are presented in Section 8.

2 EMAGE - a mouse atlas of gene expression

This paper focuses on a biological resource as its use case. That resource, EMAGE [11], publishes online gene expression information for the developmental mouse.

A gene is a unit of instructions that provides directions for one essential task, i.e., the creation of a protein. Gene expression information describes whether or not a gene is expressed (active) in a location. Such information allows biologists to discover relationships between genes, in particular when genes are active in the same location.

The gene expression information is obtained by experimenting on a mouse embryo. Each embryo corresponds to a point in time of the *developmental mouse*: the mouse from conception until birth. The time window is split into 26 distinct periods called Theiler Stages (TS). Each stage has its own anatomy, and corresponding anatomy called EMAP [10]. Moreover, there a number of 3D models representing different stages of the developmental mouse.

The result of an in-situ hybridization (ISH) experiment is documented as an image displaying an area of a mouse (from a particular TS) in which some subsections of the mouse are highly coloured, as depicted in Figure 1(A). Areas of colour indicate that the gene is expressed in that location. Furthermore, the image provides some indication of the level (strength) of expression: the more intense the colour, the stronger the expression. Results are analysed manually under a microscope. A human expert determines in which tissues the gene is expressed, and at what level of expression. Strength information is described using natural language terms such as strong, moderate, weak or present. For example, the gene *bmp4* is strongly expressed in the future brain from TS15. These statements are so-called *textual annotations*. Textual annotations represent the structured version of a subset of the original unstructured data (e.g., Figure 1(A)).

Textual annotations capture whatever information the researcher wishes to present. They may be incomplete (if the researcher is only interested in the heart, (s)he will not create textual annotations for the brain) or documented at a high granularity (the textual annotation will report the gene being expressed in the heart rather than the sub-component in which it is actually found).

In an attempt to provide a more complete and precise set of results, experimental images (e.g. Figure 1(A)) can be mapped onto 3D models of the mouse creating the *spatial annotation* depicted in Figure 1(B). These spatial annota-

tions are normally generated by EMAGE, whilst the textual annotations are produced by the researchers who performed the experiment.

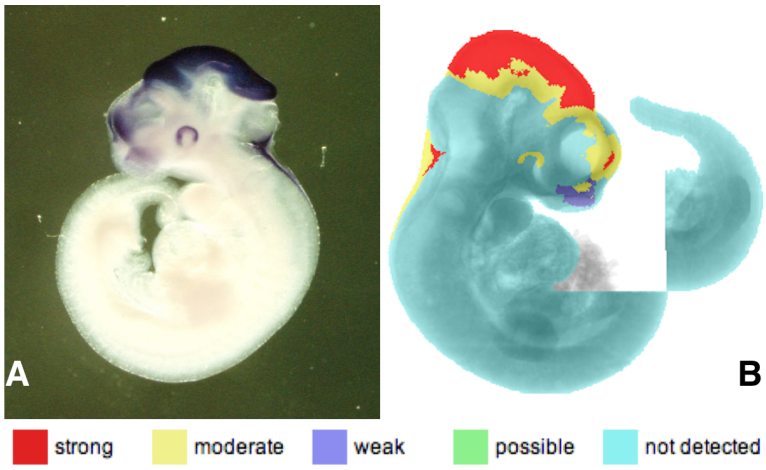

Fig. 1. A sample image of (A) an experimental result, and (B) associated spatial annotation, from the EMAGE database.

This work uses the textual annotations.

2.1 Existing EMAGE visualisations

The current range of visualisations employed within EMAGE concentrate on displaying gene expression information within the context of its location within the mouse. For example, Figure 1 (B) clearly shows where the gene *Oxt2* is expressed within TS17 (this image is taken from experiment EMAGE:1411). An alternative representation of the same information can be seen in Figure 2. In this depiction, the intensity of the colour indicates the level of expression; the greater the intensity the higher the level of expression, e.g., purple = strong and pink = moderate.

Both Figures 1 and 2 use the idea of displaying the location of gene expression on a standardised model of the mouse, with the level of expression indicated through the use of colour. Moreover, Figures 1 and 2 present the result of a single experiment (EMAGE:1411) as a spatial annotation. Accordingly, they display all the gene expression information that was obtained from the experiment. In contrast, the textual annotations report all the gene expression information that the researcher wanted to report. Currently, there is no mechanism to visualise the results of only the textual annotations. Nor is there a mechanism to represent the results (textual or spatial annotations) of multiple experiments. Instead, a user must read multiple lists of textual annotations or look at multiple spatial

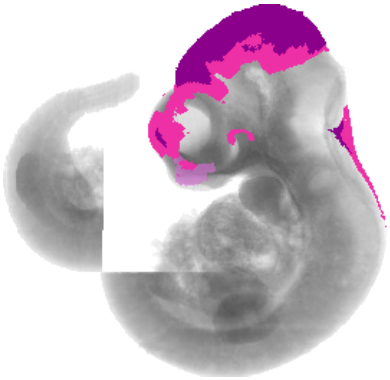

Fig. 2. An alternative representation of Figure 1 (b) in which the intensity of the colour is used to represent the level of expression, i.e., purple = strong, pink = moderate and mauve = weak.

annotation images (e.g. Figure 1(B)). Within this work, we seek to provide a visualisation tool, that enables the above tasks for the user.

EMAGE does not provide any visualisation to summarise the expression information across time or between multiple genes. For instance, it is not possible to see how the expression information for *Oxt2* compares to that of the gene *Bmp4*. Nor is it possible to see the way in which the location of *Bmp4* changes over time as the mouse develops.

3 Related work

Increasingly data-sets within the life sciences are approaching sizes which are not manageable by humans and as such usable visualisations are vital in helping human researchers navigate this data [6].

Accordingly, the life sciences are a very active area for visualisation. For example Heat Maps have been used to demonstrate Anisotropic Flocking Behaviour [1], Hive Plots [8] have been used to visualise gene co-expression and a variety of Partition Graphs [7] have been used to visualise ancestry.

Phylogenetic trees, e.g. [2], are the visualisations traditionally used to represent differences between species, and then to analyse those differences statistically.

Circos [5] was designed for the visualisation of genomic data, in particular the relationships between different cancer genomes. The CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) project uses Circos style diagrams as one mechanism for the representation of association rules. Cytoscape [9] takes this a step further, and provides a mechanism to visualise networks. This can be used to represent any biological network, for example protein interaction networks [4].

Cytoscape uses Force-directed Graphs to visualise networks. If the network can be simplified into a tree structure, Sunbust visualisations [13] or Icicle visualisations can be used instead. An icicle [15] is a sunburst transformed from polar to cartesian co-ordinates.

4 EMAGE focus group

Resources do not allow for the creation of an application with a wide array of visualisation techniques. Accordingly, it was decided to focus on a single visualisation; however, which one?

This question was answered by the EMAGE focus group: a small number of EMAGE staff who were assembled to guide this work. In the first meeting the focus group was presented with a range of different visualisation techniques and asked which they preferred. Their choice would be the mechanism featured in the application.

The favourite option was a sunburst visualisation (e.g. Figure 3; the reasons for this are simple. Firstly, the anatomy is a tree structure therefore a visualisation technique designed to display tree structures is highly appropriate. Secondly, unlike force-directed graphs (e.g., Cytoscape) there are no edges within the sunburst. This means that there are no crossed edges and no question of how to best layout the nodes. Similar results have been reported elsewhere, e.g., [13].

5 Sunburst visualisations

Essentially a sunburst (e.g., see Figure 3) takes information organised within a tree structure, and displays the tree structure in a radial layout. Assuming the information is organised as a tree, as opposed to a graph, no organisational data is lost.

The size and position of the blocks within the sunburst are used to indicate the structure, and organisation, of the data. Data attribute values are presented by colouring the nodes.

The centre of a sunburst diagram is the root node of the tree, with children of the root node being the first layer of blocks in the sunburst. Children sit directly around in the next layer of the sunburst, and so on, until the leaf nodes are reached at the edge of the diagram.

It is possible to zoom into a node by double clicking it. This causes the parent node, of the clicked node, to become the central node of the updated sunburst, and thus gives more prominence to the node of interest by making it larger. To move up a level, the user should double click on the central node. In this manner, a user may navigate up and down the internal tree structure.

6 Sunbursts for gene expression

EMAGE uses the EMAP anatomy (and corresponding ontology) to describe the anatomical space of the mouse embryo. The full anatomy is a DAG (Directed

Fig. 3. A generic sunburst diagram: each block represents a node within a tree. The order and position of the blocks recreates the structure of the tree.

Acyclic Graph); however, for computational and presentational reasons a simplified tree representation exists too. It is the EMAP tree that features within our sunburst diagrams. Because the mouse anatomy changes greatly over time, there is one sunburst for every Theiler Stage (Figure 4 shows the sunburst for TS23 with the expression profile of *Bmp4*).

Each node in the diagram represents a tissue in the mouse apart from the root node, which is the mouse itself. The node's colour is used to present the level of expression for that node, the colour scheme chosen mimics the original EMAGE colour scheme (see Figure 1).

When the user moves the mouse over a node, the box in the top right corner is updated to show the name of the tissue that node represents. Additionally, if the node contains gene expression information that is displayed.

The top left corner (Figure 4) provides a navigation box, that allows the user to move from stage to stage. In this way the user can watch the expression profile change over time. Alternatively, the same effect can be achieved by showing multiple sunbursts side by side: Figure 5 contains sunbursts for *Ssh* in stages TS14, 15 and 23.

Moreover, it is possible to show the expression profile for multiple genes in the same sunburst providing a visual way to determine which locations the genes have in common. Figure 6 shows the expression profile for over 50 genes. The genes are listed in the box in the top left corner. The coloured nodes (in the sunburst) indicate where at least one of the genes is expressed. If the mouse is moved over one of the coloured nodes the box in the top right corner is updated to show the tissue name and the list of genes expressed there (with associated

Fig. 4. Expression profile of *Bmp4* in TS23. Top left box allows navigation through other stages. Box in top right provides details of whichever tissue (node) the mouse is hovering over.

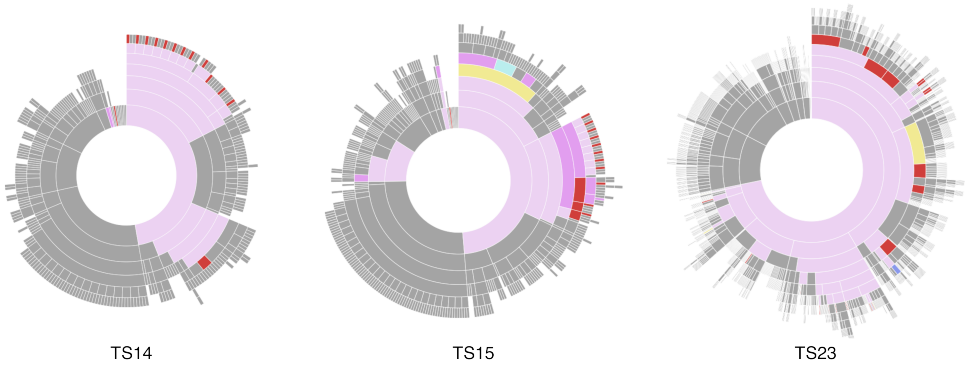

Fig. 5. Depicting the expression profile of the gene *Ssh* across time.

strengths). If multiple genes are expressed, at different strengths, in the same node the highest level of expression is used to colour the node.

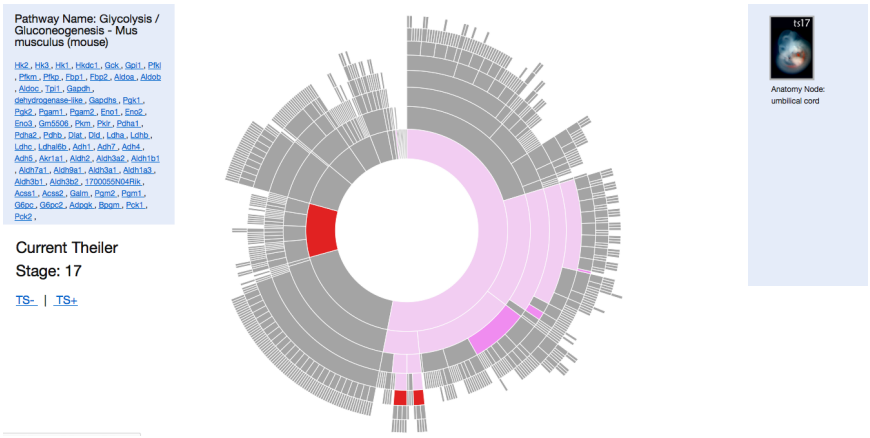

Fig. 6. Depicting the combined expression profiles of over 50 genes at TS17. Top left box lists the genes featured. Top right box shows the tissue the mouse is hovering over, and lists the genes expressed there (with associated level of expression).

7 Discussion

The basic functionality currently exists as a live system, which has been shown to the focus group. One aspect that is popular with users is the ability of the sunburst to visualise the results of multiple experiments within a single diagram. Currently, it is not possible to do this within EMAGE. One of the main disadvantages of this approach is that it only visualises textual annotations, which are often less precise/complete than spatial annotations. Moreover, it only presents textual annotations from EMAGE when it could show annotations from complementary resources (e.g., GXD [12]) too.

Whilst initial feedback was broadly positive, testing revealed some flaws with the controls: it is too difficult to change the gene(s) shown in the sunburst. Once corrected, the prototype will be shown to the focus group and their feedback incorporated. When the focus group are happy with the tool a more thorough evaluation, with a wider set of participants, will take place.

As EMAGE is one of the three CUBIST use cases it is worthwhile comparing this prototype with the CUBIST dashboard (i.e., CUBIX [3]), which also uses sunbursts. Within our work, the sunburst is used to visualise the mouse anatomy and colour indicates where a gene is expressed. In contrast, CUBIX uses sunbursts to visualise the results of Formal Concept Analysis (e.g., [14]). Rather than tissues, the nodes of the CUBIX sunburst are "concepts" and thus represent a collection of entities, for example, tissues, genes and/or Theiler Stages.

Clearly, this is early work. There is much to be considered and reviewed. For example, currently our approach to handle time (change in Theiler Stage) is to display multiple sunbursts (one for each stage), it seems impossible to do anything else. Sunburst visualisations are based on a tree structure (in this case the tree represents the anatomy at a single Theiler Stage). Whilst it would be possible to merge multiple trees by adding a new root node, this would likely lead to a diagram too complex for real world use. However, if the requirement for a sunburst is removed, it may be possible to display all the relevant information in a single diagram by switching away from a tree-based representation.

Sunburst visualisations are ideal for presenting tree structures; however, if the structure is a graph they lose information. Whilst the EMAP mouse anatomy is a directed acyclic graph (DAG) it has a simplified tree representation too. Therefore it is easy to represent the mouse anatomy as a tree, and thus sunburst. Yet, in doing so information is lost. There are many different ways of organising the mouse anatomy contained within the DAG; the "correct" way depends entirely on context. In our approach only one organisation is presented. Although this representation is the most commonly used it is not always ideal. The solution may be to offer a series of different trees/sunbursts, one for each different organisation.

Despite all the obvious problems with the sunburst, it has one attribute that is vital for this application: it is easy to understand. There is no point in creating a presentation mechanism that captures all the data and relationships if the EMAGE community find it too complex to use. A balance must be struck between presenting as much information as possible and providing a tool that users are comfortable with. Understanding where this balance lies is one of the key tasks of the current prototype.

8 Conclusion

During this paper we have presented a discussion of the ways in which sunburst visualisations can be used to present meaningful depictions of gene expression profiles. These profiles show where a gene is active, and how active that gene is. Sunburst visualisations enable a summary of the profiles to be displayed, and can present changes in the profile over time or an aggregation of multiple gene profiles. The latter is a powerful tool that enables a biologist to quickly determine what genes have in common; something that cannot be achieved with existing visualisation mechanisms in our use case. By allowing the gene and Theiler Stage to be changed, the sunbursts allow a user to visually browse the gene expression information, another feature that is currently missing from the use case.

The visualisations are being developed in partnership with biological experts from the EMAGE database of mouse gene expression. An EMAGE focus group enables us to appropriately target and test our work. Once complete, we aim to undertake a summative evaluation in order to accrue knowledge that may be applied to the development of a real world tool.

Acknowledgements

This work is part of the CUBIST project (Combining and Uniting Business Intelligence with Semantic Technologies), funded by the European Commission's 7th Framework Programme of ICT under topic 4.3: 'Intelligent Information Management'.

The authors are thankful for the advice and guidance provided by members of the EMAGE team, and for the comments provided by the anonymous reviewers.

References

1. M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1231–1237, 2008.
2. A. Boc, B. Diallo Alpha, and V. Makarenkov. T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research*, 40(W1):W5373–W579, 2012.
3. Melo C., Orphanides C., M^{c}Leod K., Aufaure M-A., Andrews S., and Burger A. A conceptual approach to gene expression analysis enhanced by visual analytics. In *Proceedings of the 20th ACM symposium on applied computing*. Coimbra, Portugal, March 2013.
4. Agapito G., Guzzi P.H., and Cannataro M. Visualization of protein interaction networks: problems and solutions. *BMC Bioinformatics*, 14(Suppl 1):S1, 2013.
5. Krzywinski M. I, Schein J.E., Birol I., Connors J., Gascoyne R., Horsman D., Jones S.J., and Marra M.A. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009.
6. Goecks J., Nekrutenko A., Taylor J., and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.
7. Baroni M., Semple C., and Steel M. A framework for representing reticulate evolution. *Annals of Combinatorics*, 8(4):391–408, 2005.
8. Krzywinski M., Birol I., Jones S., and Marra M. Hive plots - rational approach to visualizing networks. *Briefings in Bioinformatics*, 2011.
9. Cytoscape 2.8: new features for data integration and network visualization. M.E. Smoot and K. Ono and J. Ruscheinski and P.l. wang and T. Ideker. *Bioinformatics*, 27(3):431–432, 2011.
10. Baldock R and Davidson D. *Anatomy onotlogies for bioinformatics: principles and practise*, chapter The Edinburgh Mouse Atlas, pages 249–265. Springer Verlag, 2008.
11. Venkataraman S., Stevenson P., Yang Y., Richardson L., Burton N., Perry T.P., Smith P., Baldock R.A., Davidson D.R., and Christiansen J.H. EMAGE - Edinburgh Mouse Atlas of Gene Expression: 2008 update. *Nucleic Acids Research*, 36(1):D860–D865, 2007.
12. J. H. FInger, C. M. Smith, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald. The mouse gene expression database (gxd): 2011 update. *Nucleic Acids Research*, 30(suppl 1):D835–D841, 2011.

13. J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International journal of human-computer studies*, 53(5):663–694, 2000.
14. S. Andrews and K. M^{c}Leod. Gene co-expression in mouse embryo tissues. In *Proceedings of the 1dt CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) workshop*, 2011.
15. Y. Yang, P. Keller, and P. Liggesmeyer. Visual approach facilitating the importance analysis of component fault trees. In *SAFECOMP Workshops*, pages 486–497, 2012.